EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

LEAD
Graduate School & Research Network

Leibniz-Institut für
Wissensmedien

Linguistic Complexity in Longitudinal and Cross-sectional Perspectives

What characterizes the development of German-as-a-Foreign-Language learners and what is indicative of successful exam performance?

Detmar Meurers
Leibniz-Institut für Wissensmedien (IWM) &
TüCeDE, University of Tübingen

based on joint research with **Yushan Li** (Zhejiang University)

Workshop on Computational Linguistic Methods for Language Learning Technology (WRICE 2026)
Cambridge, 12./13.03.2026

Introduction

- Language tests offer snapshots of specific facets of language proficiency using psychometrically optimized test items.
  - But how can we observe linguistic competence and its development longitudinally and directly based on **contextualized, authentic language** productions?

(i) **Profile analyses** of selected linguistic means that are characteristic of specific **acquisition stages**, such as specific word orders, morphosyntax, . . .
- First language acquisition (Crystal et al.1976; Fletcher et al.2012), German (Clahsen & Hansen2012)
- Second language acquisition: German (Clahsen1985; Grießhaber2019), Rapid Profile based on Processability Theory (Pienemann1998), DAKODA project (Schwendemann et al.2025)

(ii) **Continuous characterization** based on the CAF triad (Skehan1989; Bulté et al.2025):
- CAF: **Linguistic Complexity**, Accuracy, Fluency
- “the extent to which the language produced in performing a task is **elaborate** and **varied**.” (Ellis2003)

Analysis of linguistic complexity from two different perspectives

- Linguistic complexity is being used as a measure for both **linguistic development** and for the evaluation of the **quality of texts** (Crossley2020)
  - Cf. PISA test: 30 points in test are interpreted as corresponding to learning content of one school year.
- Are the same features indicative of development and quality? (Crossley & McNamara2014)

⇒ Yushan Li and I investigated this based on corpora of Chinese learners of German.

Broad evidence for elaborateness and variety of language

- Complexity analysis traditionally reductionist: few features (Bulté & Housen2012)
  - ⇒ We investigate complexity empirically broadly across different linguistic domains (morphology, lexis, syntax, discourse) and language use (Weiss & Meurers2019a,b).

I. Which linguistic **forms of the linguistic system** occur?
   - → Theoretical linguistics, Second Language Acquisition research

II. Which **usage** of linguistic forms can be observed?
   - → Usage-based linguistics, Corpus linguistics, Psychology

III. What type & amount of **meaning** is encoded – and how organized into coherent **discourse**?
   - → Linguistics, Psychology (text comprehension)

⇒ CTAP (Chen & Meurers2016b, ctapweb.com) English: 845 features, German: 450 (Weiss & Meurers2019a,b)

Our Investigation

- Which linguistic complexity measures are characteristic for
  (a) the linguistic development of German-as-a-Foreign-Language learners
  (b) the evaluation of the quality of foreign language essays

- Data: large corpora of Chinese learners of German:
  (a) Longitudinal sub-corpus of the CDLK (Li & Wu2023)
  (b) Cross-sectional corpus PGG (PhD of Yushan Li 2025)

- Methodology:
  - analyze 450 German complexity features (Weiss & Meurers2019a,b) with CTAP (Chen & Meurers2016b)
  - data-driven selection of informative features for models using Explainable Boosting Machines (GAM)

  $\Rightarrow$ enables quantitative and qualitative analysis at different levels of granularity

Cross-sectional Corpus: Graded Essays (PGG)

- Data source: National Examination for German Majors in China (PGG)
  - 31,395 texts (5.4M tokens), 4 grades: 1-fail (7,600), 2-pass (9,769), 3-good (10,072), 4-excellent (3,954)
- nation-wide uniform exam for students of German in China for standardized assessment of learning success at the end of basic studies (end of 4th semester)
  - Data from all university types and regions, transcribed handwritten texts
  - Grading of the writing part of the high stakes PGG exam of very high quality

  → representative sample with good external validity labels (Moranski & Ziegler2021)

Longitudinal Corpus (CDLK Sub-corpus)

- Longitudinal sub-corpus of the Chinese German Learner Corpus (CDLK, Li & Wu2023).
  - 634 texts (87,766 tokens) from 4 semesters: 1 (142), 2 (172), 3 (169), 4 (151)
- Growing sub-corpus of 163 German students (3 universities from different regions)
  - Chinese native speakers with English as first foreign language
  - started learning German at university with uniform curriculum and textbook
- Four data collections at the end of each semester of the two-year basic study period
  - last collection at the time of the PGG exam
  - transcribed handwritten texts

First Step: Identification of Informative Complexity Features

- Which features are informative regarding
  - Grade evaluation in the cross-sectional corpus: 1-fail, 2-pass, 3-good, 4-excellent
  - Semester of studies in the longitudinal corpus: 1st, 2nd, 3rd, 4th semester
- Challenges:
  - many complexity features (450) and frequently high colinearity
  - Relationship complexity feature to grade/semester not necessarily linear

⇒ Explainable Boosting Machines (EBM, Lou et al.2012; Nori et al.2019)
  - Extension of Generalized Additive Models (GAM, Hastie & Tibshirani1987): machine learning (bagging, gradient boosting) for effective, iterative feature selection for model optimization

Feature Selection: Linguistic Complexity from Two Perspectives

- EBM selects from the 450 complexity features
  - 23 for Development (CDLK): Lexis 15 (55%), **Syntax 5 (22%)**, **Morph. 1 (4%)**, Discourse 2 (9%)
  - 39 for Grading (PGG): Lexis 20 (51%), **Syntax 13 (33%), Morph. 3 (8%)**, Discourse 1 (3%), Length 2 (5%)

→ all dimensions of linguistic modeling are relevant!
  - The grading perspective focuses more on syntax and morphology.

- Development and grading perspectives overlap only in 5 features:
  - Syntactic Variety (Mean Local Edit Distance for POS)
  - Syntactic Elaboration (Dependent Clause Ratio)
  - Lexical Variety (HDD excluding punctuation and numbers)
  - Lexical Frequency (SD of Verb Word Frequency per Million SUBTLEX Token)
  - and a potential Task Effect (Singular Proper Noun Density)

  but related features play a role in both (e.g., text length aspects)

⇒ Automated complexity analysis of large learner corpora enables a differentiated, variably fine-grained view on language development and text quality.

Global Characterization: Classification by Semester vs. by Grade

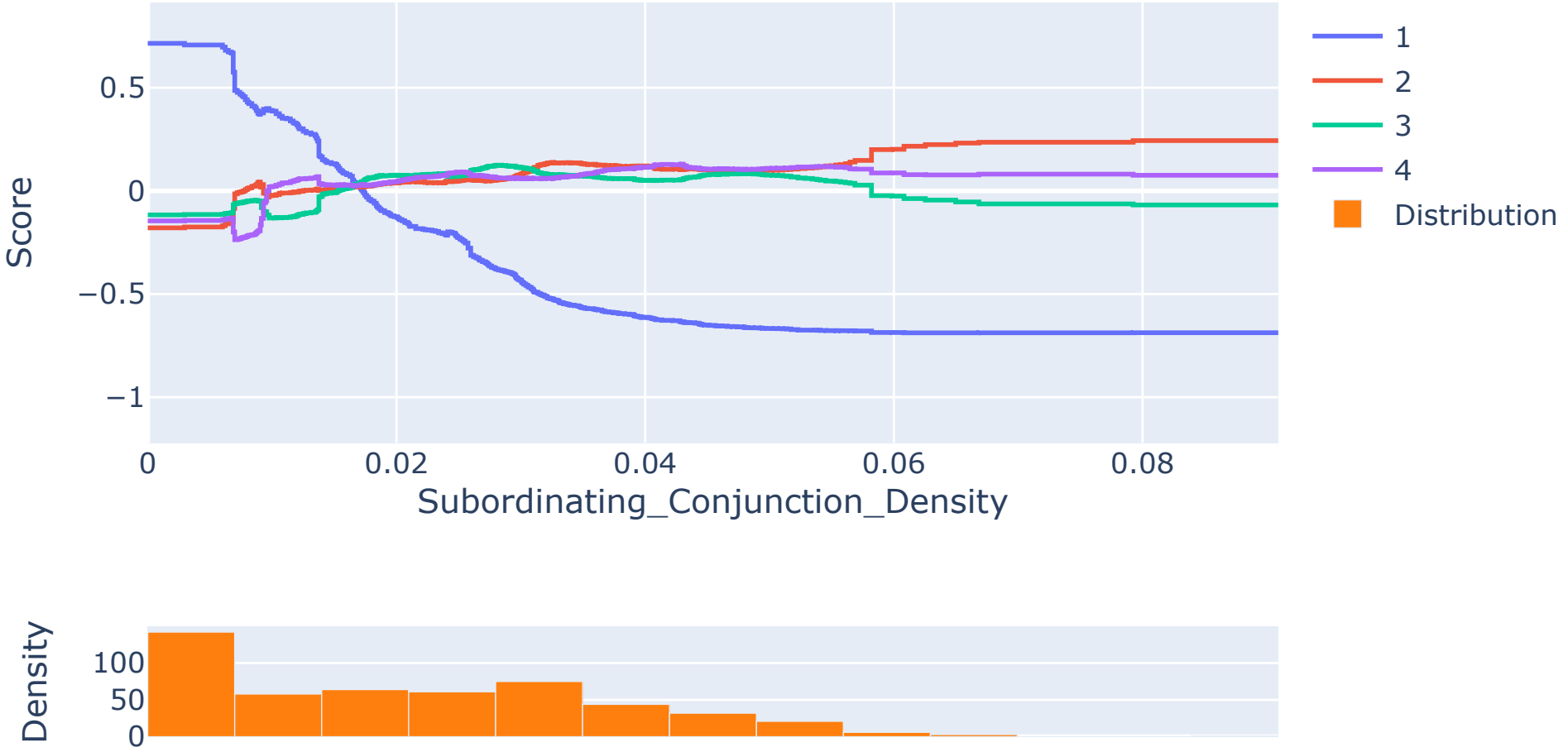
- Longitudinal: Results for classification into 4 semesters for 634 texts
  - Split: 80% Training and 20% Testing
  - Majority baseline: 27%
  - ⇒ Accuracy: 65% with 23 complexity features
- Cross-sectional: Results for evaluation with 4 grades for 31,395 texts
  - Split: 80% Training 20% Testing
  - Majority baseline: 32%
  - ⇒ Accuracy: 55% with 39 complexity features

⇒ Weaker classification for grading, although the data set is much larger.

- Grading apparently reflects other aspects besides linguistic development:
  - Accuracy
  - Appropriateness for task
  - Quality of content given the task (e.g., of argumentation)

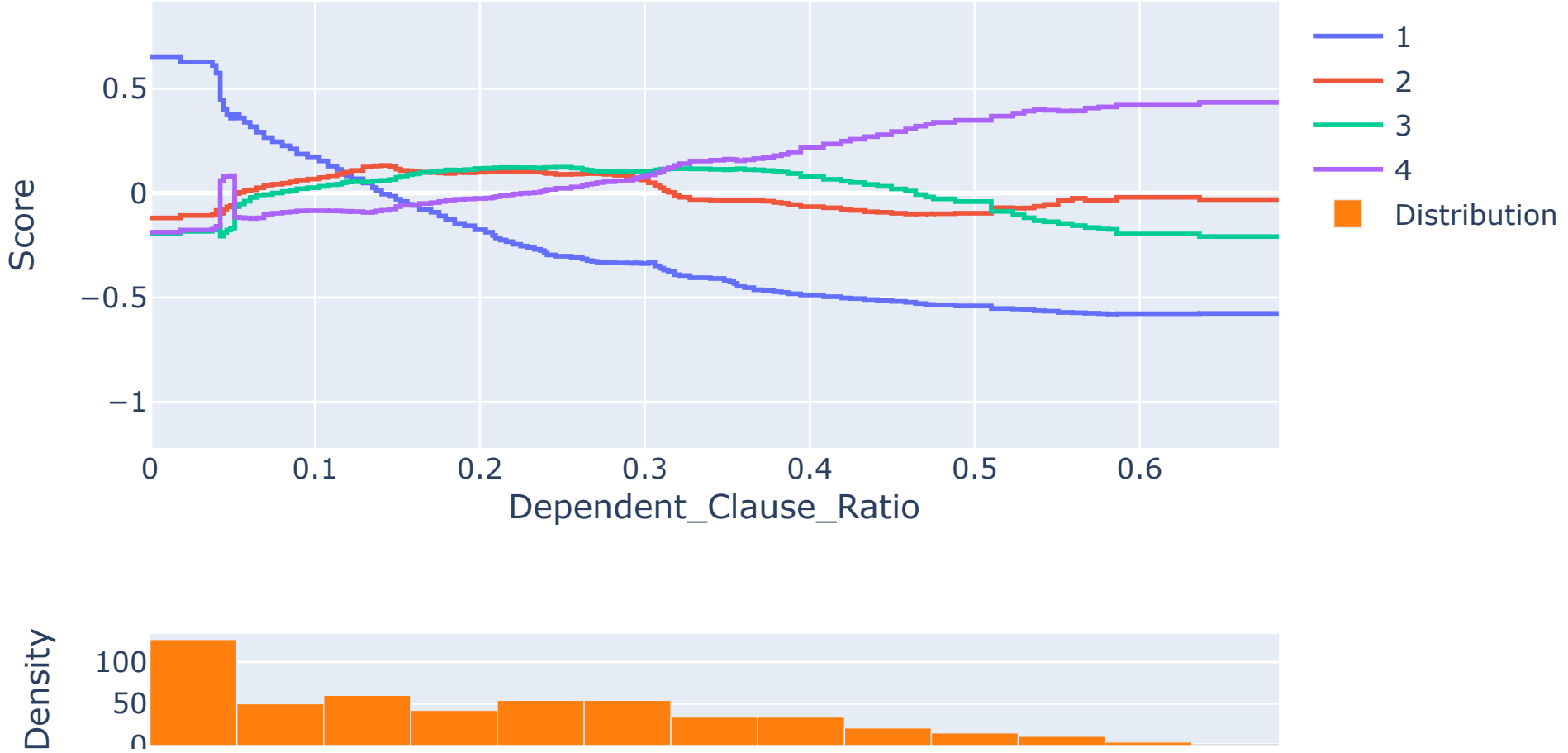
Fine-grained view on Longitudinal Development: Syntactic Elaboration

Subordinating Conjunction Density



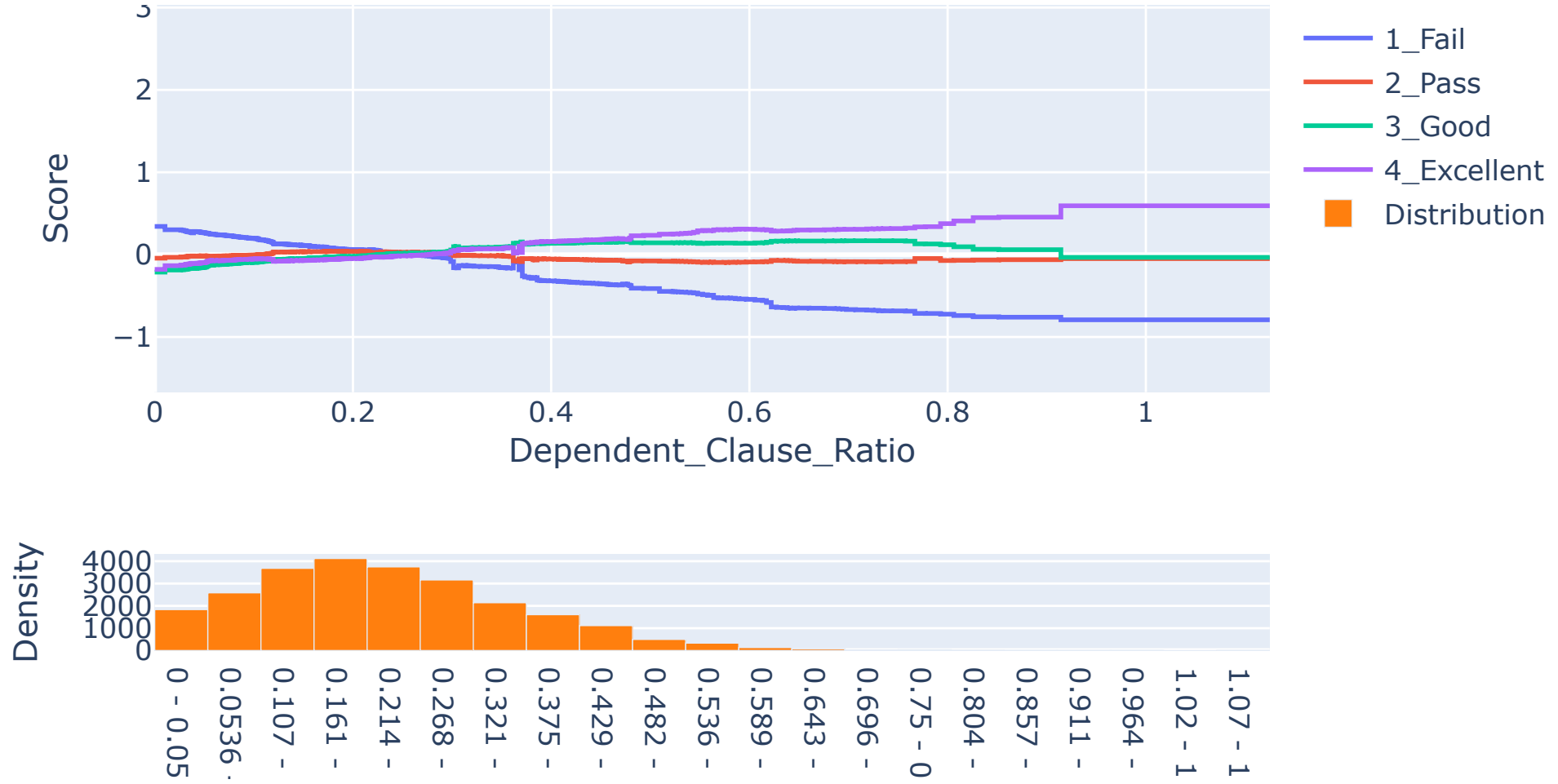
Fine-grained view on Longitudinal Development: Syntactic Elaboration

Dependent Clause Ratio



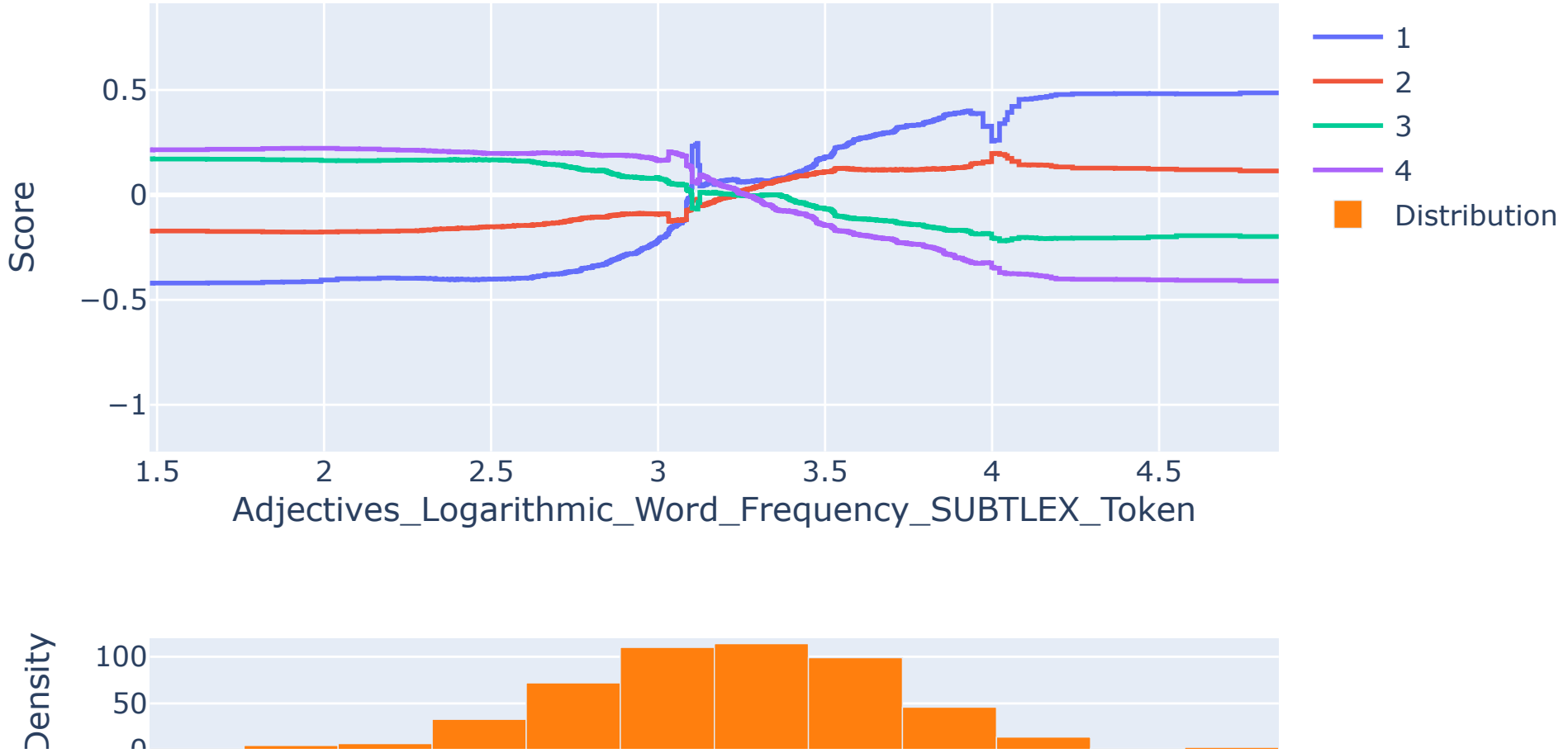
Fine-grained view on Grading: Syntactic Elaboration

Dependent Clause Ratio



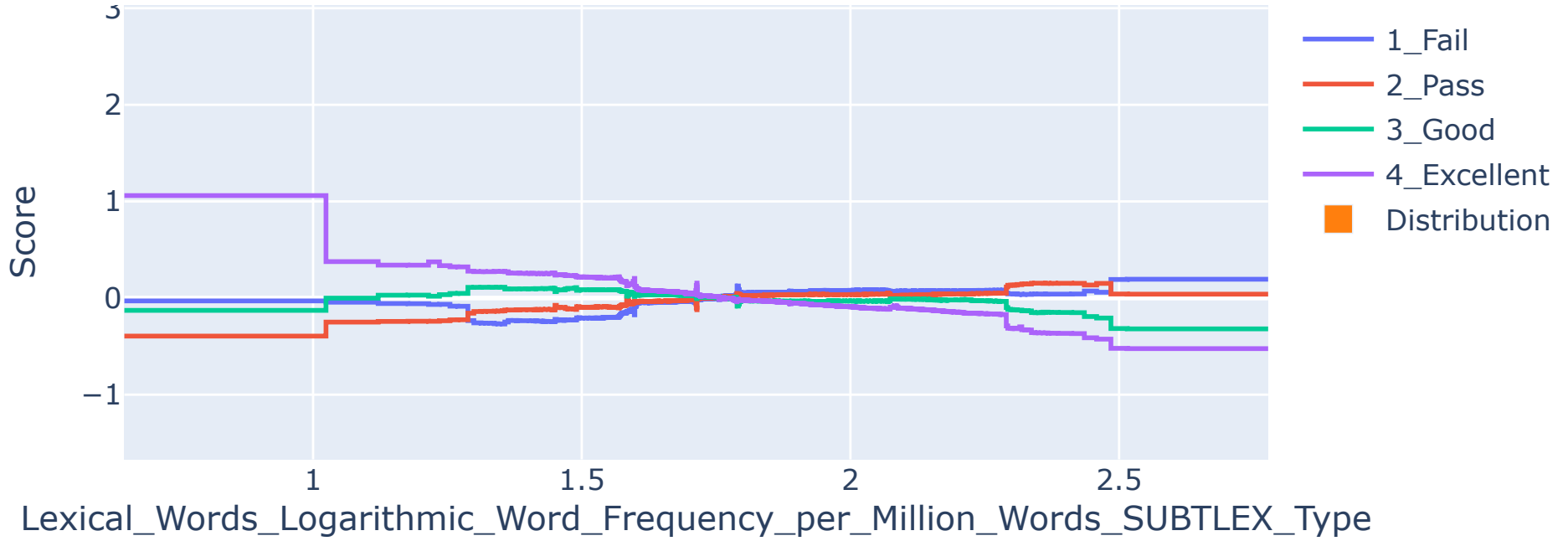
Fine-grained view on Longitudinal Development: Lexical Frequency

Adjectives Log. Word Frequency SUBTLEX Token



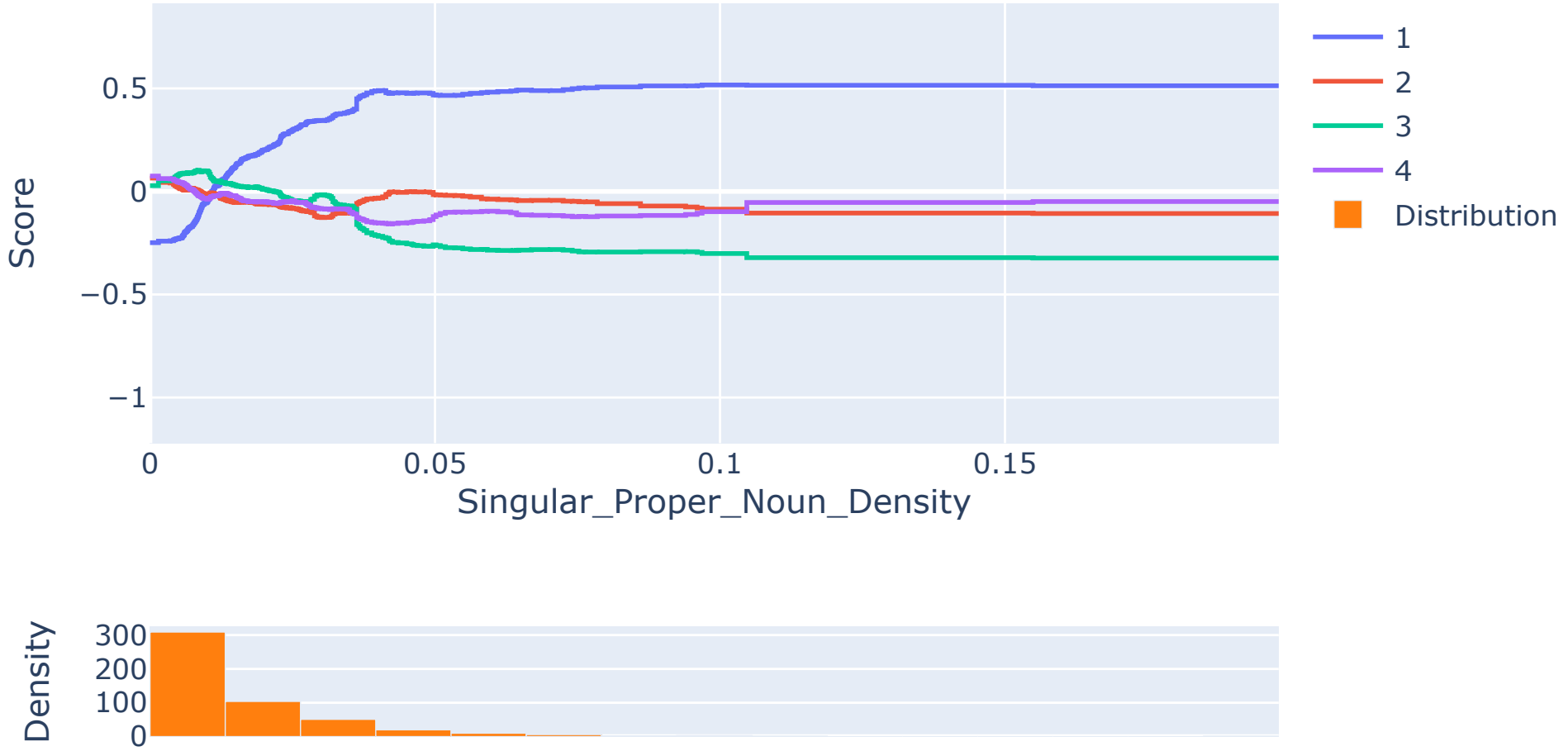
Fine-grained view on Grading: Lexical Frequency

Lexical Words Log. Word Frequency per Million Words SUBTLEX Type



Fine-grained view on Longitudinal Development: Idiosyncratic Lexical Property

Singular Proper Noun Density



Relating Task Dependence to Functional Adequacy

- Task effects (Alexopoulou et al.2017) related to the need to distinguish *appropriate complexity* (Pallotti2023) according to *functional adequacy* (Kuiken & Vedder2017)
- Complexity analysis should distinguish between
  - individually available linguistic material, and
  - its functional appropriate use.

⇒ Following a variationist linguistic perspective (Labov1972), we could distinguish:
  - Which functions are realized? (Variable)
  - Which of the linguistic means that can realize this function are used? (Variants)

Considerations on a Variationist Linguistic Perspective

- Complexity (variety, elaborateness) can be investigated on both levels:
  (i) Advanced learners can address *more varied and elaborate* ***functions*** appropriately and
  (ii) use *more varied and elaborate variants for the* ***realization*** of a function

- Example for (ii): Is functionally appropriately modification used for a picture description?
  - Are varied different variants of modification used?
    (e.g., pre-nominal/post-nominal/adverbial, varied lexical or phrasal)
  - How elaborate are the options used?
    (e.g., low-frequency adjectives, prenominal participial constructions)

From Analysis to Intervention

- Profile analyses (e.g., Language Assessment, Remediation and Screening Procedure, LARSP, Crystal et al.1976) systematically serve as basis for focused interventions.
- Unlike profile analyses, complexity analyses are not directly related to specific acquisition stages, making it more difficult to use as diagnostic measure.
- But since complexity analysis is possible for both learner output and input, interventions can provide input enrichment of developmentally proximal language!

⇒ Alignment of learners with adaptive, developmentally proximal input
  - Complex primed input continuation writing tasks (Chen & Meurers2017, 2019; Chen et al.2022)
  - Enrichment in human-computer dialogues (Glandorf & Meurers2024; Glandorf et al.2025)

Summary

- Linguistic complexity can be interpreted at different levels of granularity in relation to development or writing quality,
  - Complexity partly manifests itself in different aspects of the linguistic system, and the
  - assessment of written products is also influenced by other factors.
- Interpretation of linguistic complexity measures is dependent on the task.
  - Features are differently susceptible to task effects.
  - Task dependence observable especially for advanced learners

  → A variationist perspective on functions and their realization could advance such analyses
- Interesting options for connecting analysis to input enhancement interventions
- Outlook: Connecting Profile Analyses and Complexity Analyses
  - Extend German profile analyses with more linguistic aspects that empirically turn out to be indicative of certain stages (e.g., as found in the English Grammar Profile)?
  - Motivate complexity analyses more theoretically instead of purely data-driven.

Cited Works

Alexopoulou, T., M. Michel, A. Murakami & D. Meurers (2017). Task Effects on Linguistic Complexity and Accuracy: A Large-Scale Learner Corpus Analysis Employing Natural Language Processing Techniques. *Language Learning* 67, 181–209. URL https://doi.org/10.1111/lang.12232.

Breindl, E., A. Volodina & U. H. Waßner (2014). *Handbuch der deutschen Konnektoren 2: Semantik der deutschen Satzverknüpfer*, vol. 13. Walter de Gruyter GmbH & Co KG.

Brysbaert, M., M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte & A. Böhl (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology* 58, 412–424.

Bulté, B. & A. Housen (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (eds.), *Dimensions of L2 Performance and Proficiency*, John Benjamins, pp. 21–46. URL https://doi.org/10.1075/lllt.32.02bul.

Bulté, B., A. Housen & G. Pallotti (2025). Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning* 75(2), 533–574.

Chen, X. & D. Meurers (2016a). Characterizing Text Difficulty with Word Frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego, CA: Association for Computational Linguistics, pp. 84–94.

Chen, X. & D. Meurers (2016b). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*. Osaka, Japan: COLING, pp. 113–119.

Chen, X. & D. Meurers (2017). Challenging Learners in Their Individual Zone of Proximal Development Using Pedagogic Developmental Benchmarks of Syntactic Complexity. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017*. Gothenburg, Sweden: ACL, Linköping Electronic Conference Proceedings 134, pp. 8–17. URL http://aclweb.org/anthology/W17-0302.pdf.

Chen, X. & D. Meurers (2019). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer-Assisted Language Learning* 32(4), 418–447. URL https://doi.org/10.1080/09588221.2018.1527358. https://doi.org/10.1080/09588221.2018.1527358.

Chen, X., D. Meurers & P. Rebuschat (2022). ICALL offering individually adaptive input: Effects of complex input on L2 development. *Language Learning & Technology* 26(1), 1–21. URL https://hdl.handle.net/10125/73496.

Clahsen, H. (1985). Profiling second language development: A procedure for assessing L2 proficiency. In K. Hyltenstam & M. Pienemann (eds.), *Modelling and assessing second language acquisition*, Bristol, UK: Multilingual Matters, pp. 283–332.

Clahsen, H. & D. Hansen (2012). Profiling Linguistic Disability in German-Speaking Children. In M. J. Ball, D. Crystal & P. Fletcher (eds.), *Assessing Grammar: The Languages of LARSP*, Bristol, UK: Multilingual Matters, pp. 77–91.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research* 11(3), 415–443.

Crossley, S. A. & D. S. McNamara (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing* 26, 66–79.

Crystal, D., P. J. Fletcher & M. Garman (1976). *The grammatical analysis of language disability: A procedure for assessment and remediation*. London: Edward Arnold. Revised edition 1981.

Eisenberg, P., J. Peters, P. Gallmann, C. Fabricius-Hansen, D. Nübling, I. Barz & R. Fiehler (eds.) (2009). *Duden. Deutsche Grammatik*, vol. 4. Mannheim: Bibliographisches Institut & F. A. Brockhaus AG, 8. überarbeitete auflage ed.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford, UK: Oxford University Press.

Fletcher, P., T. Klee & W. Gavin (2012). LARSP thirty years on. In M. J. Ball, D. Crystal & P. Fletcher (eds.), *Assessing Grammar: The Languages of LARSP*, Bristol, UK: Multilingual Matters, pp. 12–28.

Glandorf, D., P. Cui, D. Meurers & M. Sachan (2025). Grammar Control in Dialogue Response Generation for Language Learning Chatbots. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics. Preprint at https://doi.org/10.48550/arXiv.2502.07544.

Glandorf, D. & D. Meurers (2024). Towards Fine-Grained Pedagogical Control over English Grammar Complexity in Educational Text Generation. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva & Z. Yuan (eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*. Mexico City, Mexico: Association for Computational Linguistics, pp. 299–308. URL https://aclanthology.org/2024.bea-1.24/.

Graesser, A. C., D. S. McNamara & J. M. Kulikowich (2012). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher* 40(5), 223–234.

Grießhaber, W. (2019). 22. Profilanalysen. In *Sprachdiagnostik Deutsch als Zweitsprache*, De Gruyter Mouton, pp. 547–568.

Hamp, B. & H. Feldweg (1997). GermaNet – a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid. URL http://aclweb.org/anthology/W97-0802.

Hastie, T. & R. Tibshirani (1987). Generalized Additive Models: Some Applications. *Journal of the American Statistical Association* 82(398), 371–386.

Keßler, J.-U. (2007). Assessing EFL-development online: A feasibility study of Rapid Profile. *Second language acquisition research: Theory-construction and testing* pp. 119–144.

Kintsch, W. (2005). An Overview of Top-Down and Bottom-Up Effects in Comprehension: The CI Perspective. *Discourse Processes A Multidisciplinary Journal* 39(2 & 3), 125–128.

Kuiken, F. & I. Vedder (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing* 34(3), 321–336.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Li, Y. & Z. Wu (2023). Chinesisches Deutschlerner-Korpus (CDLK). Ein umfangreiches Korpus mit Mehrebenen-Annotationen und multidimensionalen Metadaten. In M. Kupietz & T. Schmidt (eds.), *Neue Entwicklungen in der Korpuslandschaft der Germanistik*, Tübingen: Gunter Narr Verlag, pp. 223–236.

Lou, Y., R. Caruana & J. Gehrke (2012). Intelligible Models for Classification and Regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, KDD '12, pp. 150–158.

Mackey, A., M. Pienemann & I. Thornton (1991). Rapid Profile: A second language screening procedure. *Working Papers of the National Languages Institute of Australia* 1(1), 61–82.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Ph.D. thesis, University of Memphis. URL https://umdrive.memphis.edu/pmmccrth/public/Papers/MTLD%20dissertation.doc.

McCarthy, P. M. & S. Jarvis (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42(2), 381–392.

Moranski, K. & N. Ziegler (2021). A case for multisite second language acquisition research: Challenges, risks, and rewards. *Language Learning* 71(1), 204–242.

Nori, H., S. Jenkins, P. Koch & R. Caruana (2019). InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* URL https://arxiv.org/abs/1909.09223.

Pallotti, G. (2023). Appropriate complexity. In C. Granget, I. Repiso & G. F. Sing (eds.), *Language, creoles, varieties: From emergence to transmission*, Berlin: Language Science Press, pp. 145–173.

Pienemann, M. (1998). *Language Processing and Second Language Development: Processability Theory*. Amsterdam: John

Benjamins.

Schwendemann, M., K. Wisniewski, L. Lenort, A. Portmann, C. Renker, J. Ruppenhofer & T. Zesch (2025). Zur Entstehung und Erschließung der bislang größten Lernerkorpus-Datenbank des Deutschen: Ein Bericht aus dem DAKODA-Projekt. *Zeitschrift für Germanistische Linguistik* 53(3), 580–600.

Skehan, P. (1989). *Individual Differences in Second Language Learning*. Edward Arnold.

Weiss, Z. & D. Meurers (2019a). Analyzing Linguistic Complexity and Accuracy in Academic Language Development of German across Elementary and Secondary School. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. Florence, Italy: Association for Computational Linguistics.

Weiss, Z. & D. Meurers (2019b). Broad Linguistic Modeling is Beneficial for German L2 Proficiency Assessment. In A. Abel, A. Glaznieks, V. Lyding & L. Nicolas (eds.), *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*. Louvain-La-Neuve: Presses Universitaires de Louvain.

Lexicon

- Lexical **Diversity** (*say* vs. *explain, assert, . . .*)
  - Type-Token Ratio = *Typ*/*Tok*
  - Text length independent: Measure of Textual Lexical Diversity (MTLD, McCarthy2005)
- Lexical **Frequency** (*Ship* more frequently used than *barque*)
  - Word frequency in SUBTLEX-DE (Brysbaert et al.2011)
  - When aggregating, also consider variance and clustering methods (Chen & Meurers2016a)
- Lexical **Semantics**
  - Synonymy: Number of words with same meaning (GermaNet Synsets, Hamp & Feldweg1997)
  - Polysemy/Homonymy: Number of different word senses of a word

Morphology

- **Derivation**
  - Nominalization (*Zerstör-ung (destruction), Heiter-keit (cheerfulness)*, . . .)
  - Compounds (*Stadt-entwicklung-s-potential* (urban development potential), . . .)
- **Inflection**
  - Case (Genitive, . . .)
  - Verb forms (Subjunctive, . . .)

Syntax

- Systematic analysis of sentences, T-Units, clauses:
  - Mean **Length** (average sentence length, . . .)
  - **Number** of occurrences (Number of clauses per sentence, . . .)
  - Occurrence of **subtypes** (subordinate clauses, . . .)
- **Structural Variety**
  e.g., Parse edit distance between adjacent sentences, or globally in text
- Number and length of selected **elaborated constructions**
  e.g., complex noun phrases per sentence

Discourse

- Connectives (temporal, causal, . . .)
  - lexically based on Breindl et al.2014) and Eisenberg et al.2009)
  - challenging to determine whether an occurrence connects propositions
- Cohesion based on repetition of language material (local, global, cf. Graesser et al.2012)

Topics in CDLK and PGG Corpora

- Longitudinal CDLK sub-corpus: *My family*, *An unforgettable experience*, *Image description*, *Work or Master's degree?*, *Should cell phones be allowed in the classroom?*
- PGG cross-sectional corpus:

Year	Topic	# Texts	# Tokens
2012	*Interns wanted for Chinese classes*	2.783	494.525
2013	*Writing to parents to convince them to agree to their child's participation in a rural mission*	983	206.536
2014	*Father and son*	2.233	481.430
2015	*Parents must limit children's computer time*	3.164	570.865
2016	*Emigration*	3.285	586.975
2017	*Do a Master's or work straight away*	3.371	537.032
2018	*The five most popular subjects of study*	3.432	550.801
2019	*Proportion of men with overweight and obesity in Germany in the years 2005 to 2017*	4.195	714.894
2020	*Debt among young people*	3.968	666.475
2021	*Time management*	3.981	616.631
		31.395	**5.426.164**

Longitudinal Corpus: CDLK Sample of 163 Learners

Group	Semester 1		Semester 2		Semester 3		Semester 4		Sum	
	Text	Token	Text	Token	Text	Token	Text	Token	Text	Token
1	27	3.520	29	4.147	27	3.528	24	3.207	107	14.402
2	26	2.672	24	3.072	23	3.600	20	2.832	93	12.176
3	23	2.239	23	3.116	24	3.330	23	3.379	93	12.064
4	36	3.427	37	4.925	36	4-605	36	4.192	145	17.149
5	30	4.997	59	8.121	59	10.025	48	8.832	196	31.975
Sum	**142**	**16.855**	**172**	**23.381**	**169**	**25.088**	**151**	**22.442**	**634**	**87.766**

Analysis of Classification by Semester

Semester	Precision	Recall	F1-Score	Support
1	**93%**	**93%**	93%	28
2	62%	74%	68%	35
3	55%	**47%**	51%	34
4	**50%**	**47%**	48%	30

- **Very good identification of beginners**
- **Weaker differentiation of advanced learners**
- **Especially also lower Recall**

Analysis of Classification by Grade

	Precision	Recall	F1-Score	Support
1-Fail	73%	63%	68%	1520
2-Pass	49%	49%	49%	1954
3-Good	50%	65%	56%	2014
4-Excellent	55%	**26%**	36%	791

- **Weak Recall for excellent essays**
  - supports hypothesis that grading considers other aspects than linguistic development